APPARATUS AND METHOD FOR SPLIT TRANSISTOR MEMORY HAVING IMPROVED ENDURANCE

TECHNICAL FIELD

The present invention relates generally to semiconductor memory devices, and in particular to floating gate transistor structures used in non-volatile semiconductor memory devices such as flash memory devices.

BACKGROUND OF THE INVENTION

Flash memory devices are high density, non-volatile memory devices having low power consumption, fast access times and low cost. Flash memory devices are thus well suited for use in a variety of portable electronic devices that require high density storage but cannot support a disk drive, or other mass storage devices due to high power consumption or the additional weight of such devices. An additional advantage of flash memory is that it offers in-circuit programmability. A flash memory device may thus be reprogrammed under software control while the device resides on a circuit board within an electronic device.

Figure 1 is a flash memory cell 10 according to the prior art. The flash memory cell 10 has a metal oxide semiconductor (MOS) structure that includes a substrate 12, a pair of source/drain regions 14, a floating gate 18 overlying a MOS channel region 16, and a control gate 20 overlying the floating gate 18. An oxide structure 22 separates the floating gate 18 from the channel region 16, and also separates the floating gate 18 from the control gate 20. For the device shown, the substrate 12 is doped with P-type impurities, and the source/drain regions 14 are doped with N-type impurities.

The memory cell 10 may be programmed by applying a sufficiently positive gate voltage V_{CG} and a positive drain voltage V_D to the device 10, while maintaining the source voltage V_S at a zero, or ground potential. As charge is moved to the floating gate 18 from the source/drain region 14, the device 10 attains a logic state "0". Alternately, if little or no charge is present at the floating gate 18, a logic state corresponding to "1" is stored on the device 10.

To read the state of the device 10, a positive voltage V_{CG} of predetermined magnitude is applied to the control gate 18, while V_D is maintained positive. If the voltage applied to the control gate 18 is sufficient to turn the device 10 on, a current flows from one source/drain region 14 to the other source/drain region 14 that may be detected by other external circuits, thus indicating the logic state "1". Correspondingly, if sufficient charge exists at the floating gate 18 to prevent the device 10 from turning on, a logic state of "0" is read. A logic state may be erased from the device 10 by applying a positive source voltage V_S to the source/drain region 14 while V_{CG} is maintained at a negative potential. The device 10 attains a logic state "1" following an erase cycle.

Although the foregoing flash memory cell 10 is highly effective to store a logic state in a memory device, it has been observed that the programming efficiency of the memory cell 10 is degraded as the number of accumulated program/erase cycles increases. As a result, the cell 10 may fail after the number of program/erase cycles exceeds a limiting value, which is termed the endurance limit for the cell 10. Although the endurance limit is relatively unimportant in cases where the cell 10 is programmed only once, it may be a critical concern where the device 10 is erased and reprogrammed numerous times. The degradation of the programming efficiency is believed to result from hot electrons that become trapped in the relatively thin oxide layer separating the floating gate 18 from the substrate 12 during a programming cycle, which permanently damages the oxide layer. In addition, extremely high electric field strengths are generated during erase cycles that cause holes having relatively low momentum to become trapped in the oxide layer separating the floating gate 18 and the substrate 12. As the cell 10 is subjected to repeated program/erase cycles, the trapped holes accumulate in the oxide layer and thus cause the electric fields applied during a read cycle to be degraded.

The qualitative effects of degradation of the flash memory cell 10 are shown in Figures 2-4. Figure 2 compares the performance of a non-cycled flash memory cell 10 with the performance of the cell 10 after it has been subjected to a substantial number of erase and programming cycles. As shown in Figure 2, the source/drain current I_{DS} for the cycled cell 10 is significantly lower that that obtained from a non-cycled cell 10 for a

comparable fixed control gate voltage V_{CG}. As a consequence, the determination of a logic state during a read cycle is adversely affected due to the lowered source/drain current in the cycled cell 10. This effect is further shown to Figure 3, where the source/drain current I_{DS} of the cell 10 is observed to steadily decrease as the number of cycles accumulates on the cell 10. Figure 3 also shows that the endurance limit for the cell 10 may occur between approximately 10^5 and 10^6 cycles.

Figure 4 shows the variation of a threshold voltage V_T for the cell 10 as the number of program/erase cycles is increased. The threshold voltage V_T is defined as the minimum required voltage to turn on a cell 10 during a read cycle. In Figure 4, $V_{T,1}$ corresponds the threshold value required to turn on the cell 10 when the floating gate of the cell 10 is charged (indicating logic state "0"), while $V_{T,2}$ corresponds to the threshold value required to turn on the cell 10 when the floating gate 18 is not charged. The difference between the $V_{T,1}$ and $V_{T,2}$ values thus defines a threshold voltage "window", as shown in Figure 4. As the cell 10 is subjected to cycling, the "window" becomes progressively smaller, so that it becomes more difficult to distinguish between the two logic states stored in the cell 10.

One prior art solution to the foregoing endurance limit problem is a flash memory cell having a floating gate asymmetrically positioned towards the source, with the control gate overlying the floating gate and also directly overlying the channel region of the cell, as disclosed in detail in an article by P. Pavan, *et al.*, entitled "Flash Memories-An Overview", *IEEE Proceedings*, vol. 85, No. 8, pp. 1248-1271, 1997. Since the programming and erase functions occur in the portion of the channel region adjacent to the source, damage to the gate oxide is limited to only a portion of the channel region. Although the foregoing flash memory cell arrangement achieves some increase in the endurance limit, the damage to the oxide layer underlying the floating gate eventually becomes excessive, so that it is no longer possible to read the logic state stored in the cell.

Another prior art flash memory cell includes a source region that is surrounded by an N- region to further protect the source junction of the cell from the large electric field strengths that arise when the cell is erased. One significant drawback present

in this configuration is that the source and drain regions may not be interchanged to extend the endurance of the cell. Further, the asymmetrical arrangement adds to the overall fabrication costs of the flash memory device.

Accordingly, there is a need in the art for a flash memory device having an enhanced endurance limit.

SUMMARY OF THE INVENTION

The present invention is directed towards systems, apparatuses and methods for forming floating gate transistor structures used in non-volatile semiconductor memory devices such as flash memory devices. In one aspect, the system may include a central processing unit (CPU), and a memory device coupled to the processor that includes an array having memory cells, each cell including a first columnar structure and a spaced apart second columnar structure having a floating gate structure interposed between the first columnar structure and the second columnar structure and spaced apart from the first and second structures, the floating gate being positioned closer to a selected one of the first and second structures. In another aspect, a memory device includes an array having memory cells having first and second adjacent field effect transistors (FETs) having respective source/drain regions and a common floating gate structure that is spaced apart from the source/drain regions of the first FET by a first distance, and spaced apart from the source/drain regions of the second FET by a second distance. In still another aspect of the invention, a method of forming a memory device having a plurality of interconnected memory cells includes positioning a first columnar structure on a substrate, positioning a second columnar structure on the substrate that is spaced apart from the first columnar structure, forming a gate structure between the first structure and the second structure; and interposing a floating gate structure between the first structure and the gate structure and between the second structure and the gate structure, the floating gate structure being positioned closer to selected one of the first structure and the second structure.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a cross sectional view of a flash memory cell according to the prior art.

Figure 2 is a graph that qualitatively compares the drain/source current performance for a cycled and a non-cycled flash memory cell.

Figure 3 is graph that qualitatively illustrates the degradation of the drain/source current performance as the number of cycles is increased for a flash memory cell.

Figure 4 is graph that qualitatively illustrates the narrowing of the voltage threshold window of a flash memory cell as the number of cycles is increased.

Figure 5 is a block diagram of a computer system 100 according to an embodiment of the invention.

Figure 6 is a block diagram of a memory device according to another embodiment of the present invention.

Figure 7 is a partial schematic diagram of a memory cell array according to an embodiment of the invention.

Figure 8 is a partial isometric view of a portion of a memory cell array according to an embodiment of the invention.

Figure 9 is a partial cross sectional view of a memory array according to an embodiment of the invention.

Figure 10 is a partial plan view of a memory array according to an embodiment of the invention.

Figure 11 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 12 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 13 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 14 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 15 is a partial plan view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 16 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

Figure 17 is a partial cross sectional view that illustrates a step in a method for forming a memory array according to another embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is generally directed to semiconductor memory devices, and in particular to floating gate transistor structures used in non-volatile semiconductor memory devices such as flash memory devices. Many of the specific details of certain embodiments of the invention are set forth in the following description and in Figures 5-17 to provide a thorough understanding of such embodiments. One skilled in the art will understand, however, that the present invention may be practiced without several of the details described in the following description. Moreover, in the description that follows, it is understood that the figures related to the various embodiments are not to be interpreted as conveying any specific or relative physical dimension. Instead, it is understood that specific or relative dimensions related to the embodiments, if stated, are not to be considered limiting unless the claims expressly state otherwise.

Figure 5 shows an embodiment of a computer system 100 that may use the memory device of Figures 6-17 or some other embodiment of a memory device according to the present invention. The computer system 100 includes a processor 102 for performing various computing functions, such as executing specific software to perform specific calculations or tasks. The processor 102 includes a processor bus 104 that normally includes an address bus, a control bus, and a data bus. The processor bus 104 is coupled to a memory controller 106, which is, in turn, coupled to a number of other components. The

processor 102 is also typically coupled through the processor bus 104 to a cache memory 107, which is usually a static random access memory ("SRAM") device.

The memory controller 106 is coupled to system memory in the form of a synchronous random access memory ("SDRAM") device 108 through an address bus 110 and a control bus 112. An external data bus 113 of the SDRAM device 108 is coupled to the data bus of the processor 102, either directly or through the memory controller 106.

The memory controller 106 is also coupled to one or more input devices 114, such as a keyboard or a mouse, to allow an operator to interface with the computer system 100. Typically, the computer system 100 also includes one or more output devices 116 coupled to the processor 102 through the memory controller 106, such output devices typically being a printer or a video terminal. One or more data storage devices 118 are also typically coupled to the processor 102 through the memory controller 106 to store data or retrieve data from external storage media (not shown). Examples of typical storage devices 118 include hard and floppy disks, tape cassettes, and compact disk read-only memories (CD-ROMs). Finally, the memory controller 106 is coupled to a basic input-output ("BIOS") read only memory ("ROM") device 120 for storing a BIOS program that is executed by the processor 102 at power-up. The processor 102 may execute the processor 102 either directly from the BIOS ROM device 120 or from the SDRAM device 108 after the BIOS program has been shadowed by transferring it from the BIOS ROM device 120 to the SDRAM device 108. The BIOS ROM device 120 is preferably a non-volatile memory device according to the present invention, such as the embodiments of the invention shown in the memory device of Figures 6-17. Memory devices according to present embodiments may also be used in the computer system 100 for other functions.

Figure 6 is a block diagram of a memory device 200 according to an embodiment of the present invention, which may comprise at least a portion of the memory 108 shown in Figure 5. The memory device 200 includes a memory cell array 210 that includes memory cells comprised of floating gate FET transistor devices as will be described in greater detail below. The memory device 200 also includes an x-gate decoder 230 that provides a plurality of gate lines XG1, XG2...XGN for addressing the cells in the

memory cell array 210. A y-source/drain decoder 240 provides a plurality of source/drain lines YD1, YD2...YDN for accessing the first source/drain regions of the floating gate FET transistor cells in the array 210. An x-source/drain decoder 250 similarly provides a plurality of data lines XS1, XS2...XSN for accessing second source/drain regions of the cells in the memory array 210. The x-source/drain decoder 250 may also include sense amplifiers and input/output (I/O) devices for reading, writing or erasing data from the memory cell array 210. The memory device 200 further includes address buffers 220 that receive address signals A0...AN from the address bus 140 (as shown in Figure 5). The address buffers 220 are coupled to the x-gate decoder 230, the y-source/drain decoder 240 and the x-source/drain decoder 250 to control the reading, writing and erasing operations on the memory cells in the memory cell array 210.

Figure 7 is a partial schematic diagram illustrating an embodiment of the memory cell array 210, as shown in Figure 6. The memory cell array 210 includes a plurality of adjacent and interconnected memory cells 300 of substantially similar configuration that extend in a first direction along a row of the array 210 from a cell 300AA to a cell 300AN. The array further extends in a second direction to a row 300 NA that further extends in the first direction to a cell 300NN. Each of the memory cells 300AA through 300NN includes a pair of field effect transistors (FETs) 310 having an electrically isolated floating gate that controls the conduction between the source and drain regions in the FETs 310. The FETs 310 in each of the cells 300AA to 300 NN share a common gate, such as XG1, XG2....XGN, and are formed in columnar structures, as described in greater detail below.

Figure 8 is a partial isometric view illustrating a portion of the memory cell array 210 of Figure 7. For clarity of illustration, only memory cells 300AA and 300AB of the array 210 are shown, and in the following description, only memory cell 300AA will be described. It is understood, however, that the array 210 includes a substantial number of cells having a substantially similar structure, so that the array 210 extends in a first direction (the "x" direction, as shown in Figure 8), and also in a second direction (the "y" direction, also as shown in Figure 8) that is substantially perpendicular to the first direction.

The cell 300AA includes a pair of columnar structures 328A and 328B formed on a p-type substrate 320. Each of the columnar structures 328 includes a first source/drain region 322 comprised of a material having an N+ conductivity that extends along the substrate 320 in the x-direction. The structures 328A and 328B further include a second source/drain region 326 also having an N+ conductivity that is positioned adjacent to the first source/drain region 322. A separation layer 324 of material doped to have a conductivity of P- is interposed between the first source/drain region 322 and the second source/drain region 328.

Still referring to Figure 8, the columnar structures 328A and 328B are spaced apart to permit the gate line XG1 to be positioned between the structures 328A and 328B. A floating gate 330 is interposed between the structure 328A and the gate line XG1, and between the structure 328B and the gate line XG1. The floating gate 330 further extends below the gate line XG1 so that the floating gate 330 is also interposed between the gate line XG1 and the underlying substrate 320 to form a single control gate 330 between the structures 328A and 328B. The floating gate 330 is electrically isolated from the gate line XG1 by a first dielectric layer 340 that is interposed between the gate line XG1 and the floating gate 330. The floating gate 330 is further electrically isolated from the first structure 328A and the second structure 328B by a second dielectric layer 350 interposed between the floating gate 330 and the structures 328A and 328B. The floating gate 330 is further positioned between the first structure 328A and the second structure 328B so that the floating gate 330 is positioned closer to the first structure 328A than to the second structure 328B, as will be shown in greater detail below. Accordingly, a portion of the second dielectric 350 that is substantially adjacent to the first structure 328A is thinner than a corresponding portion of the second dielectric 350 that is adjacent to the second structure 328B. One skilled in the art will recognize, however, that the thinner portion of the second dielectric 350 may be positioned adjacent to the second structure 328B, while a thicker portion of the second dielectric 350 is positioned adjacent to the first structure 328A. The floating gate 330 may be comprised of a polysilicon material that is deposited on the array 210 during a fabrication process, as will also be described in greater detail below. The first

dielectric layer 340 and the second dielectric layer 350 may be comprised of silicon dioxide that is grown or deposited during the fabrication of the array 210, although other similar dielectric materials may also be used.

The second source/drain region 326A of the first structure 328A and the second source/drain region 326B of the second structure 328B are interconnected by a data line YD1 that is comprised of a metallic or other interconnection line that is substantially electrically isolated from the underlying topology of the array 210. Accordingly, it is understood that the array 210 as shown in Figure 8 may be overlaid by a layer of a dielectric material (not shown) that includes contact penetrations that are etched in the dielectric material in order to permit the data line YD1 to be connected to the first structure 328A and the second structure 328B.

Figure 9 is a partial cross sectional view of the memory array 210 that is viewed from the section line 9-9 of Figure 8, and thus viewed generally parallel to the x-direction shown in Figure 8. As noted above, the floating gate 330 is separated from the first structure 328A and the second structure 328B by dissimilar thicknesses of the second dielectric layer 350. Accordingly, the first structure 328A is spaced apart from the floating gate 330 by a first distance d_1, and the second structure 328B is spaced apart from the floating gate 330 by a second distance d_2, where the first distance d_1 is less than the second distance d_2. In a particular embodiment, the second distance d_2 is approximately about two times the thickness of the first distance d_1. In another particular embodiment, the floating gate 330 has a height d_3 of approximately about 0.1 μm, and is spaced apart from the first and second structures 328A and 328B by a first distance d_1 of approximately about 33 Å and a second distance d_2 of approximately about 66 Å.

Figure 10 is a partial plan view of the memory array 210 shown in Figure 9. In particular, the cell 300AA has a pitch that extends in the y-direction of approximately about 2F, and a pitch that extends in the x-direction approximately about 2F, where F is characteristic dimension associated with a minimum lithographic feature size. Accordingly, a logic state corresponding to a single data bit may be advantageously stored

in an area of approximately about $4F^2$. This compares favorably with a feature size of $8F^2$ for the well-known folded array architecture commonly found in DRAM memory arrays.

The foregoing embodiment provides still other advantages over the prior art. For example, and with reference again to Figure 9, since programming and erase functions are performed on the first structure 328A that is spaced apart from the floating gate 330 by a generally thinner portion of the dielectric layer 350, charge trapping in the thinner oxide layer will have only a minor effect on the opposing second structure 328B that is positioned adjacent to a generally thicker portion of the dielectric layer 350 during read operations.

Figures 11-16 are partial cross sectional views that illustrate steps in a method for forming a memory array according to another embodiment of the invention. Referring first to Figure 11, a substrate 320 formed from silicon and doped to a P- conductivity is used as a starting material. A first source/drain region 322 is formed on the substrate 320. The region 322 may be formed on the substrate 320 by ion implantation or other similar processes in order to attain the desired N+ conductivity. Alternately, an epitaxial layer of N+ silicon may be grown on a surface of the substrate 320. A separation layer 324 may then be formed on the first source/drain region 322 by an epitaxial growth of P- silicon to a desired thickness. A second source/drain layer 326 may be formed on the separation layer 324 by another epitaxial growth of N+ silicon. A pad layer 400 comprised of silicon oxide may be formed on an exposed surface of the second source/drain layer 326, which may be overlayed by a pad layer 420, comprised of silicon nitride.

Turning now to Figure 12, a plurality of first trenches 440 and a plurality of second trenches 460 are formed in the structure shown in Figure 11. The first trenches 440 and the second trenches 460 are formed in the structure of Figure 11 in a direction that is approximately perpendicular to the y-direction and are further substantially mutually parallel. The first trenches 440 and the second trenches 460 project downwardly into the structure to the p-subtrate layer 320. The first trenches 440 and the second trenches 460 may be formed by patterning an exposed surface of the structure shown in Figure 11 with a layer of photoresist (not shown in Figure 12) to form an etch barrier having exposed surface portions that coincide with the intended locations of the first trenches 440 and the second

trenches 460. The substrate material underlying the exposed surface portions may be removed by plasma etch methods, or by wet etching method known in the art.

Still referring to Figure 12, the first trenches 440 and the second trenches 460 are substantially filled with silicon dioxide 480 that is grown in the first trenches 440 and second trenches 460 through an oxidation process, or deposited in the first trenches 440 and second trenches 460 by other well-known methods. The material positioned between the first trenches 440 and the second trenches 460 (as shown in Figure 12) is removed by forming another etch stop layer of photoresist (not shown) and removing the material by wet or plasma etch methods to form voids 500, as shown in Figure 13. A bottom portion 510 comprising a silicon dioxide material is formed by oxidation, or other well-known deposition processes to form the second dielectric layer 350.

Referring now to Figure 14, a polysilicon layer 520 is formed on the structure of Figure 13, which extends downwardly into each of the voids 500 of Figure 13. The polysilicon layer 520 may be deposited on the structure by various well-known methods. An oxide layer 530 is then formed on the polysilicon layer 520 by exposing the polysilicon layer 520 to an oxidation process. A polysilicon or metal layer 540 may then be formed over the oxide layer 530 by various well-known polysilicon or metal deposition methods.

Figure 15 is a partial plan view that illustrates the formation of a plurality of substantially parallel grooves 520 that extend in the y-direction. The grooves 520 are formed by selectively etching the structure shown in Figure 14, so that the polysilicon or metallic interconnections 530 extend across the grooves 520. The interconnections 530 form the gate lines XG1, XG2...XGN as described in detail in connection with Figures 8-10. The polysilicon layer 520, the oxide layer 530 and the polysilicon or metal layer 540 may then be removed from the upper surfaces 540, as shown in greater detail in Figure 16. The layers 520, 530 and 540 may be removed using chemical-mechanical planarization.

Turning to Figure 17, a surface oxide layer 550 may be deposited on a surface 550 and patterned using a photoresist (not shown) to form an etch-stop layer to

form a plurality of protrusions 590 that extend through the surface oxide layer 550 to the second source/drain regions 326. A metal layer 570 is then deposited on the surface oxide layer 550 that extends downwardly into each of the protrusions 590 to electrically couple the second source/drain regions 326, forming the data lines YD1, YD2...YDN described in detail in connection with Figures 8-10.

From the foregoing it will be appreciated that, although specific embodiments of the invention have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of the invention. For example, certain features shown in the context of one embodiment of the invention may be incorporated into other embodiments as well. Accordingly, the invention is not limited by the foregoing description of embodiments except as by the following claims.